Open Source Science for ESO Mission Processing Study

Identify a system architecture that meets the ESO mission processing objectives, supports open science, enables system efficiencies, and promotes earth-system science.

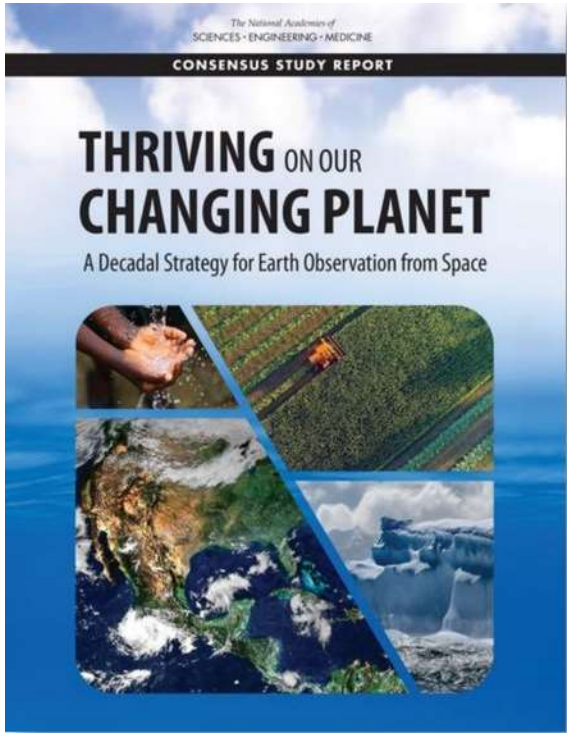
Workshop #2
March 1-4, 2022

Earth Information System (EIS) Perspective

Alexey N. Shiklomanov

NASA Goddard Space Flight Center, Biospheric Sciences Lab (618)

On behalf of the EIS Team



"While some discoveries are grounded entirely on observations from space, many more depend on **combining information from a range of sources**, including **field campaigns, laboratory experiments, computer modeling, and theoretical studies** [...].

Science based on **integrating information from several approaches** can lead to products where the **insights from the whole are much greater than the sum of the parts**."

NAS Earth Science 2017-2027 Decadal Survey

pg. 2-23

Earth Science Decadal Survey questions need clear, actionable answers

(W-5) What processes determine the spatial and temporal patterns of air pollutants?

(H-1) How is the water cycle changing? And how do these changes affect the frequency and magnitude of extremes such as droughts and floods?

(S-3) How will local sea level change along coastlines around the world in the next decade to century?

EIS pilot studies — March-September 2021 (6 months!)

Fire

Harnessing NASA's unique data and models to understand the impacts of new extreme fires in the earth system

Freshwater

Integrating data and models across the full water cycle to deliver actionable freshwater information

Sea level change

Advancing understanding of sea level change by breaking barriers to collaboration and connecting process models to observations

Common themes

(1) Tackled complex Earth Science Decadal Survey questions by synthesizing state-of-the-art models and observations from NASA and its partners.

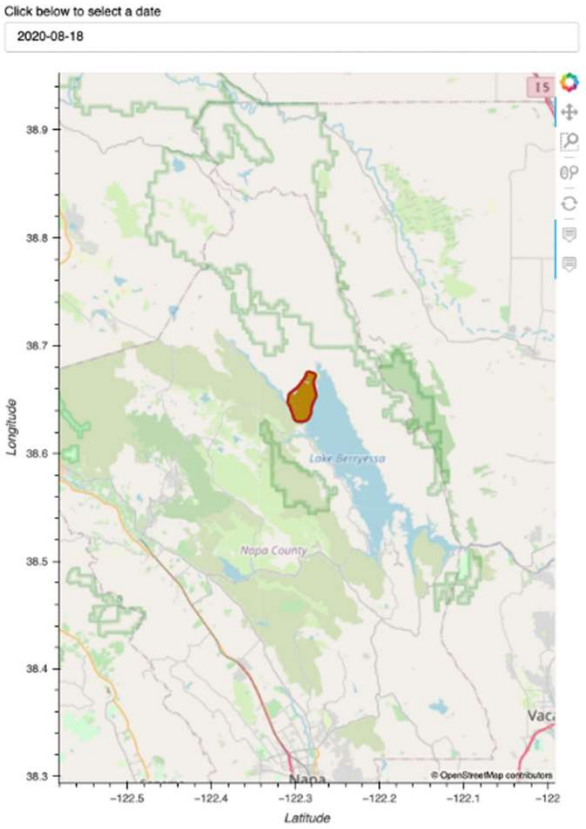
(2) Facilitated interdisciplinary scientific collaboration and stakeholder engagement leveraging open-source tools and emerging computing capabilities

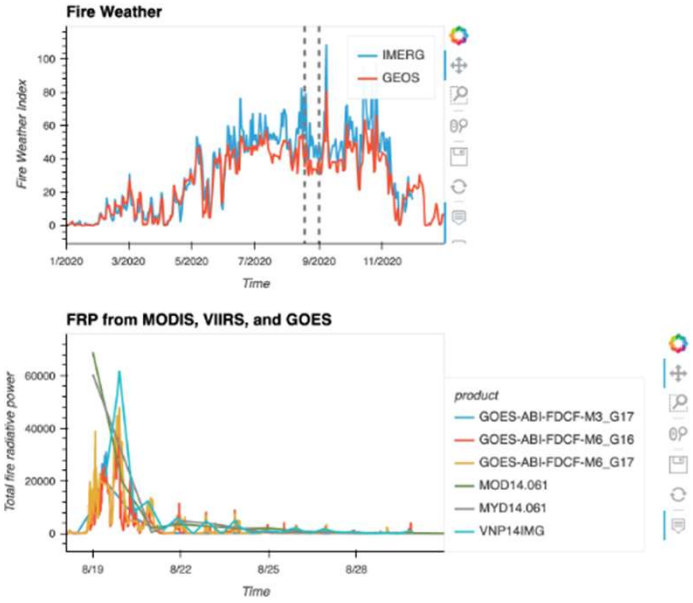
(3) Translated scientific results into actionable information for a wide range of users and stakeholders.

EIS-Fire

NASA fire data is scattered across many data portals and distribution mechanisms. **EIS-Fire** *brought this information to a single location, allowing people to comprehensively study individual fires.*

- Demonstrated the ability to **track individual fires using VIIRS data**
- Assimilated VIIRS active fire data to **improve fire emissions in GEOS**.
- Developed interactive dashboards, powered by open community standards and tools, to respond directly to stakeholder needs.

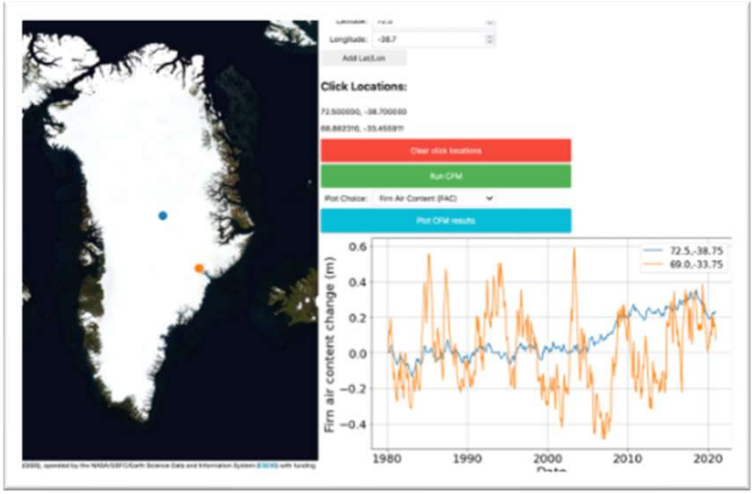




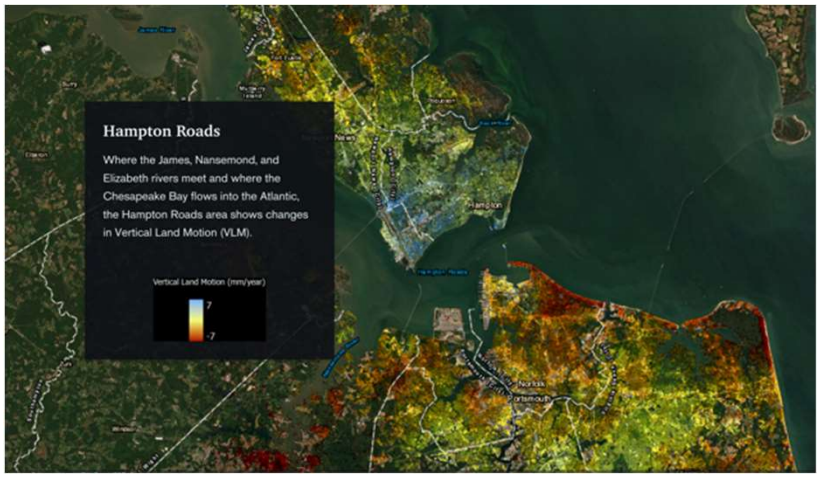
This example shows a time series of multiple satellite and model datasets spatially averaged over the selected fire perimeter.

EIS-Sea Level

Evaluated the contribution of the Greenland Ice Sheet to sea-level change and its local impacts



Graphical user interface (powered by Jupyter) for a process-based firn model

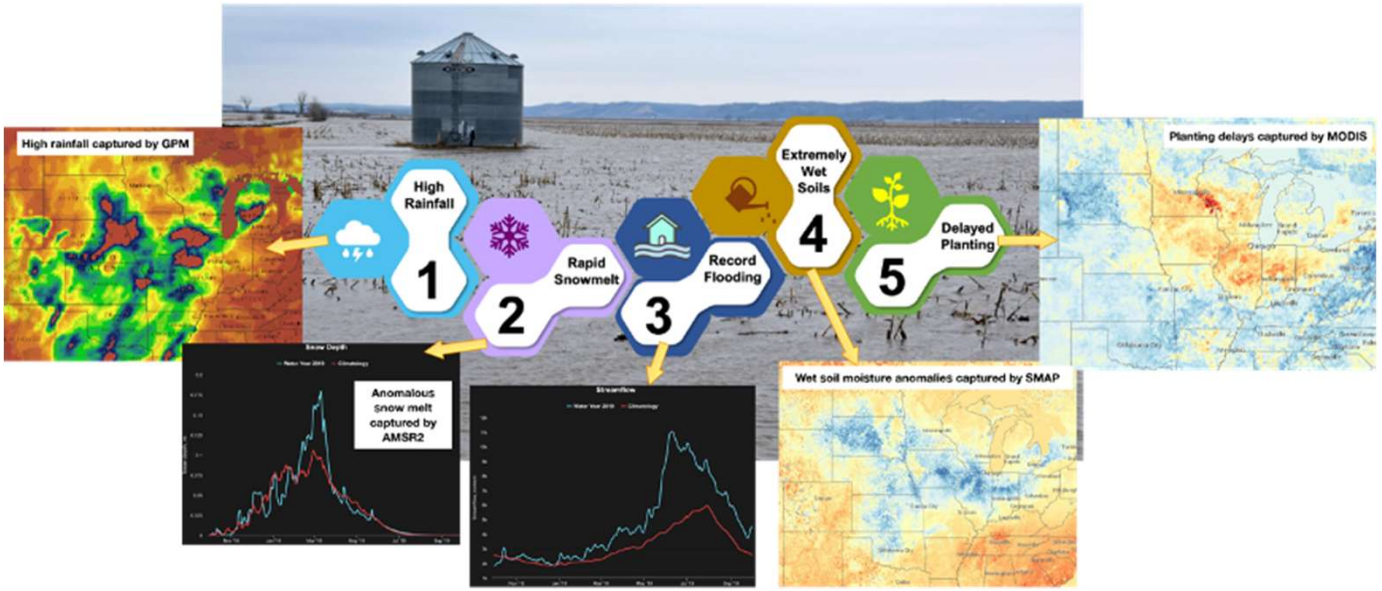


Map of vertical land motion in Hampton Roads, VA, visualized via ESRI StoryMap.

- **Implemented four process models on the *Science Managed Cloud Environment (SMCE)*** that simulate various cryospheric processes.
- Combined models with satellite observations of altimetry, mass change, and related variables to **map coastal flooding risk at the neighborhood level**
- **Leveraging open science principles** allowed users to easily share code, reproduce results, and build on each other's work.

EIS-Freshwater

Integrating data and models across the **full water cycle** to deliver **actionable freshwater information**



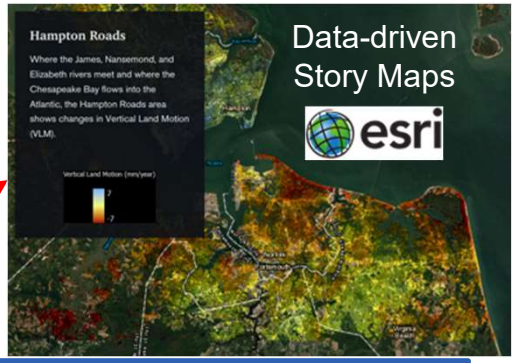
Case study: 2019 floods in the US Midwest — A perfect storm of compounding factors: high rainfall, rapid snowmelt, flooding, drenched soils, and ultimately delayed planting.

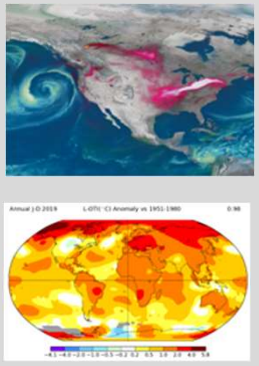
- Captured key processes and their impacts by including **multiple constraints from NASA remote sensing instruments**—GPM, SMAP, GRACE-FO, AMSR2, and MODIS—within the models.
- Open science environment (SMCE) enabled **rapid prototyping of models**, including remote sensing-informed hydrology outputs of LIS used to develop water quality outputs with SWAT.
- **Effectively described water cycle science simply**, appealing to a broad set of users

Project Goals and Objectives

- What are the key and driving requirements?
  - **Analysis-ready access** to many different NASA and non-NASA data products, including ground-based, airborne, and satellite observations and model-derived products
  - Ability to run simulations of **computationally intensive physically-based models**
  - **Interactive development and collaboration environment** allowing analysis-in-place, to which new users (including non-NASA users) can be easily added
  - **Publicly-accessible interfaces and APIs** for subsetting, analysis, and visualization
- Who are the customers and stakeholders?
  - **Scientists** — easier to do new research and develop new products, including iterating on existing products
  - **Operational agencies and decisionmakers**— easier to use NASA models, data, and expertise to support operations and decisions
  - **General public** — making NASA science activities more accessible
- What external factors constrain your solution?
  - Fragmentation of NASA's Earth Science capabilities
  - NASA IT security policies
  - NASA copyright and intellectual property rules, and the NASA Software Release process
  - Competitive, small-project funding model for NASA that discourages collaboration

System Architecture





Earth System models

Open-source tools and standards

PANGEO DASK HoloViz xarray

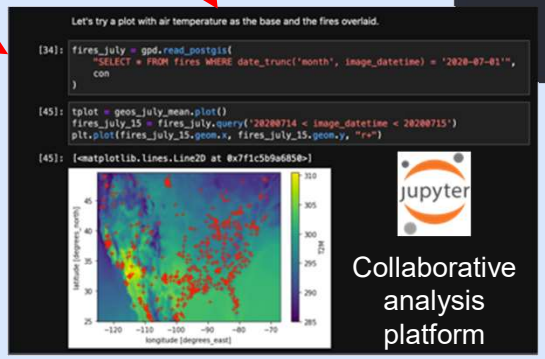
amazon S3

Cloud-optimized storage

Zarr

Flexible, on-demand compute

NASA Land Information System LIS



jupyter

Collaborative analysis platform

Science Managed Cloud Environment (SMCE)

Component and Infrastructure View

- **Science Managed Cloud Environment (SMCE)** – low-security NASA entry point to AWS — is great for rapid development and external collaboration.
- **Modular design**, re-using open-source components wherever possible
- **Infrastructure as code** – component stack can be re-deployed at the push of a button.

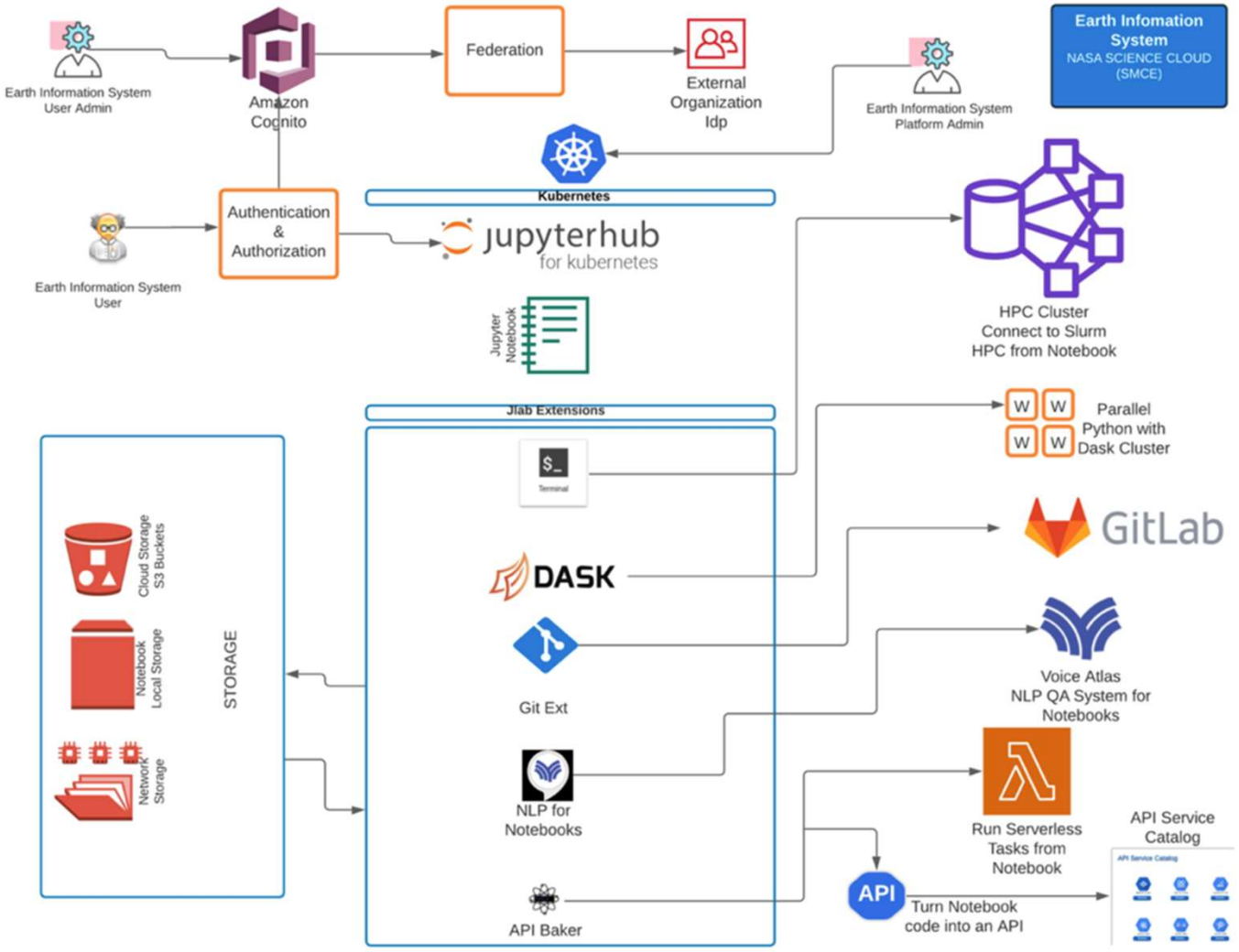


Diagram credit: SMCE team

Implementation, Deployment, and Operations

- Does the project reuse legacy software?
  - **Yes!** Many simulation models are written in **Fortran or C** and don't lend themselves well to (traditional) Jupyter notebook execution.
- Does the project use open-source software?
  - **Yes! Pangeo stack**, **open-source geospatial tools** (including from Pangeo community)
- Are software developed in open and collaborative environments?
  - **Yes!** Shared Jupyter Notebooks, version control via self-hosted GitLab.
  - Straightforward process for granting new users access
- How does the project handle cybersecurity?
  - Everything assumed to be **FISMA Low**.
  - JupyterLab access controlled through **AWS Cognito**. Cluster access controlled through **SSH keys**.
  - Relatively straightforward process for granting users AWS Console Access.

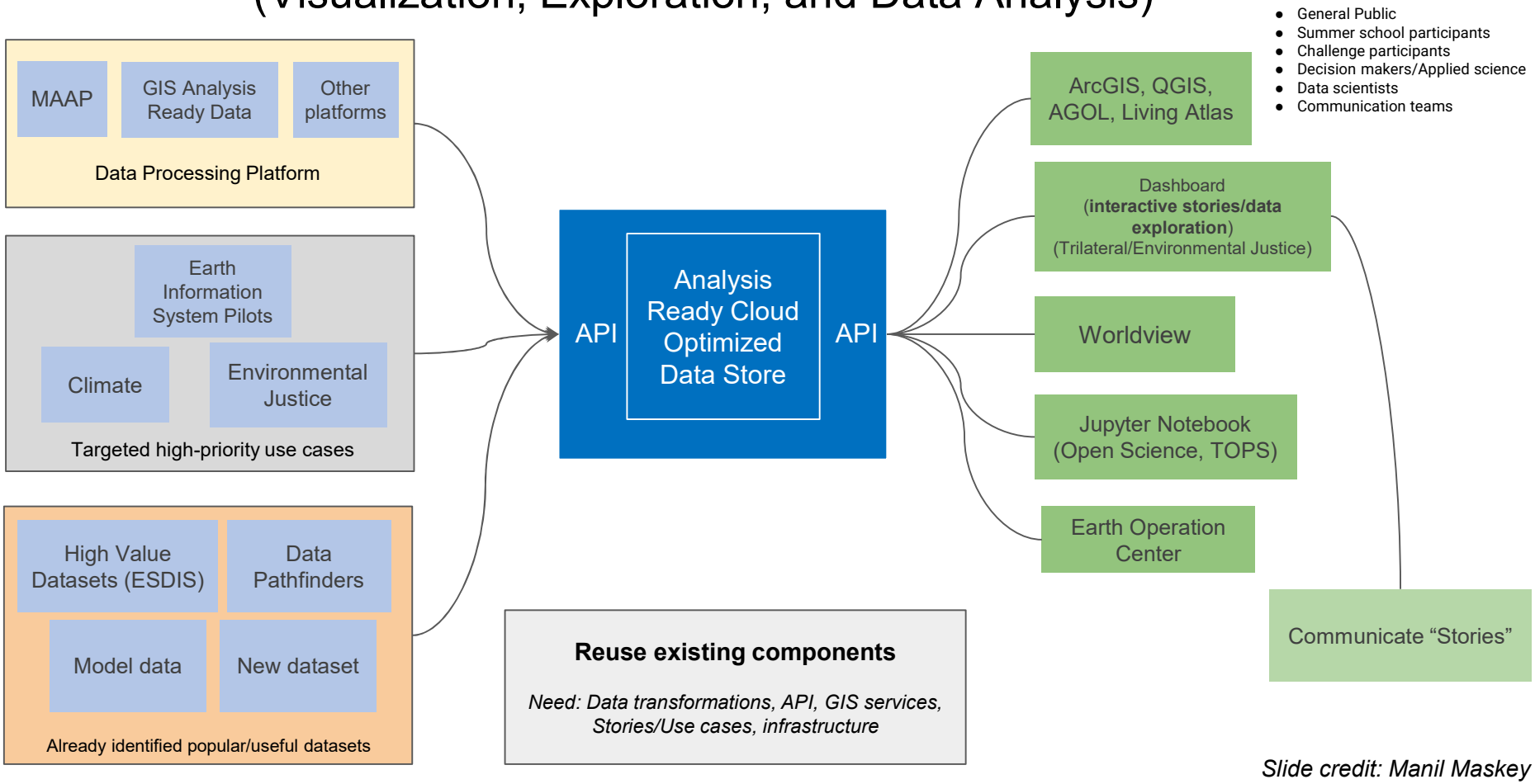
Implementation, Deployment, and Operations (continued)

- What compute environments does the system use?
  - **Science Managed Cloud Environment (SMCE)**: low-security, development-oriented NASA access point to **AWS**
- How has the project ensured system efficiency (cost, storage, processing time, etc.)?
  - SMCE “comes with” a **system administration support team** that monitored costs (using AWS CloudWatch)
  - Aggressive use of **EC2 spot instances**
- How do you deploy and operate your system?
  - **SMCE system administration team** was excellent. This is an essential component and is not free!
  - **Infrastructure as code** made it easy to track software stack

Open Science and Community Collaboration

- What does open science mean to your project?
  - **Open = Transparent** — Code was released in publicly available GitLab repositories.
  - **Open = Inclusive** — Collaborators from other Federal and Local government agencies and universities used our platform to do new science.
  - **Open = Reproducible** — Infrastructure as code; conda environments; Jupyter notebooks
  - **Open = Accessible** — Translating complex NASA observations and model outputs into actionable information and compelling stories
- What barriers have your project faced?
  - **Short project timeline with frequent deliverables** — Tension between producing quick wins and building a robust system.
  - **NASA Software Release Process** — NASA has no institutional support for a reasonable open-source development model, and a long and difficult process for releasing open-source code at the end of a project.
- How has your project enabled community collaboration?
  - We invited people to use our stuff, granted them access, and allowed them to contribute…
  - …but we did not contribute to other NASA projects or external open-source communities.

What's next for EIS? VEDA!
(Visualization, Exploration, and Data Analysis)



Slide credit: Manil Maskey

Lessons learned

- Co-locate data in a small number of common, analysis-ready formats
  - NOTE: There are no silver bullets for analysis-ready storage — need to evaluate trade-offs based on most common use cases
- Don't reinvent the wheel! Take advantage of community standards and software ecosystems.
- Use a versatile system that is allowed to fail fast and has low security requirements
- Invest in software development and system administration support — this is not cheap!
- Engage users and stakeholders early and often — they are good at telling you what they need and can help keep you on track.
- **Stakeholders don’t want products, they want information, solutions, and answers**! We need to move beyond mission-specific thinking.